Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Environmental Sciences 14 (2012) 247 – 256

Procedia Environmental Sciences

Landscape, Environment, European Identity, 4-6 November, 2011, Bucharest

Applications of Principal Component Analysis integrated with GIS

Alexandru-Ionuț Petrişor[a,b,*], Ioan Ianoş[c], Daniela Iurea[d], Maria-Nataşa Văidianu[c]

[a]*Department of Urban and Landscape Planning, School of Urbanism, "Ion Mincu" University of Architecture and Urbanism, Bucharest 010014, Romania*

[b]*National Institute for Research and Development in Constructions, Urbanism and Sustainable Spatial Development URBAN-INCERC, Bucharest 021652, Romania*

[c]*Interdisciplinary Center for Advanced Researches on Territorial Dynamics, University of Bucharest, Bucharest 030018, Romania*

[d]*Faculty of Geography, University of Bucharest, Bucharest 010041, Romania*

Abstract

Principal Component Analysis is a statistical instrument able to identify the variables explaining most variation within a sample. When the lines are administrative units within a region, and the input variable account for a specific issue (e.g., level of development etc.), Principal Component Analysis can be used to pinpoint the variables explaining mostly the specific issue. Moreover, if used in conjunction with GIS modeling, the entire approach produces hierarchies of the administrative units, which by mapping allow for the identification of 'hotspots' (e.g., underdeveloped regions etc.) that are at the core of intervention policies. The presentation examines several examples from research aiming to assess the level of development in Romania, Ialomița hydrographic basin, the Danube Delta Biosphere Reserve and Iaşi County. The results indicate the utility of the approach as a research instrument, but also for strategic planning purposes. Specifically, they allow for pinpointing the variables that best account for the level of development, describing economic, social, demographic, education, infrastructure and cultural aspects. The results indicate that development cannot be assessed using variables pertaining to a single sector, as in all cases the relevant variables account for economic, social, cultural, or demographic issues.

Keywords: geostatistics; territorial systems; spatial development; variability

* Corresponding author. Tel.: +4-021-307-7133; fax: +4-021-312-3954.
E-mail address: alexandru.petrisor@incd.ro

1878-0296 doi:10.1016/j.proenv.2012.03.024

1. Introduction

One of the definitions of geography calls it the "science of the organized space" [1]. The organization process is the effect of the activity of human communities, as a consequence of the tendency of man to arrange objects surrounding him [2]. At a different spatial scale, territorial organization is made based on the socio-economic targets of human communities [2]. The organizational units are called territorial systems [3], [4], and defined as "functional assemblies constituted of elements and relations, aiming to achieve common goals" [5]. Based on their size, territorial systems can be assimilated to the levels of the Nomenclature of Units for Territorial Statistics (NUTS) hierarchy [6][6, [7].

The study of territorial systems has been carried out traditionally using an arsenal of methods having as main common features their subjectivity, conducing to the lack of reproducibility of results [8] and reduced ability of quantification. The importance of the first shortcoming has been stressed out in the literature for a long time [9]. To overcome subjectivity, one of the possible solutions is importing methodologies from other sciences [5, 8, 10]. Of particular importance are the statistical methods; their advantage is the ability to quantify the degree of uncertainty when generalizing the results of individual case studies [11], making the results comparable and reproducible [12].

The application of statistical methods to specific fields has given birth to new disciplines due to their specific adaptations to the data requirements and interpretation of the results specific to the discipline using them [13]. For this reason, this paper proposes expanding the meaning of the term "geostatistical methods", currently restricted to spatial prediction techniques such as kriging [14] to include all methods situated at the interference of geography and statistics. Petrişor [8] proposed a hierarchy of these methods based on the degree of mathematical abstractness and research objectives, ranging from purely qualitative geographical methods tightly connected to the territorial reality to abstract statistical and mathematical models loosing their geographical relevance.

A dilemma in using statistical methods in geography is that geography involved and still involves, at least in some countries, a tight link with specific places and an interpretation of results based on their particularities. In contrast, statistics (especially inferential) tends to generalize the results [13]. The use of statistical methods assumes sampling, determination of the characteristics of each individual unit within the sample, and the generalization of these features to the entire population [15].

To exemplify, while for biologists it is not hard to generalize the characteristics of some fish forming a sample to their entire species [16] and epidemiologists can easily ascribe the features of subjects with a certain condition to all people exhibiting it relying on statistical methods [17], it seems to be very hard for geographers to see beyond the features of a given city and regard it as a data line for a particular typology.

Nevertheless, there is a situation when generalization is easily grasped even by geographers. This particular case relates to the organization of territories. For statistical and administrative purposes, there are territorial divisions corresponding to the levels of the Nomenclature of Units for Territorial Statistics (NUTS) hierarchy. A territorial unit can be seen as a sample composed of units represented by lower NUTS levels. In this setting, different branches of geography could be concerned with specific research questions, but the goal is essentially common: to produce a hierarchy of the units and analyze its spatial distribution. To achieve this goal, geographers rely on data derived from territorial statistics on as many variables as possible. However, not all variables accounted for are of equal importance in explaining the spatial variability. To determine the most influential variables for a given territorial system and use them to look at the spatial distribution of their values in the subunits of the system, this paper proposes an approach combining two methods, one statistical (Principal Component Analysis) and one geographical (Geographical Information Systems modeling). The first method results into the determination of the most influential variables and their weights based on the percentage of variability explained. These outputs are then embedded in a Geographical Information System model to produce and map a hierarchy of spatial

subunits based on the values of the most influential variables for each subunit. A similar approach, combining GIS and PCA, was used in recent studies in ecology [18], while PCA was used to investigate different geographical issues in Romania [19, 20, 21, 22, 23, 24, 25].

The aim of this paper is to present several examples documenting the application of the method at several spatial scales, and outline its utility in identifying the variables that most accurately describe the level of development within the subunits of a given space, and pinpoint the most developed or undeveloped units, in an attempt to prove that the level of development cannot be accurately described by variables looking at a single aspect (e.g., economic, social, cultural or environmental).

2. Methodology

As stated previously, the approached proposed by this paper combines Principal Component Analysis (PCA) and Geographical Information Systems (GIS) modeling. PCA is a statistical method aiming for the reduction of data, identifying components that account for the overall variability within the variables taken into consideration; the principal components are linear combinations of these variables accounting for the common and unique variability explained by them [26].

The steps involved by the application of PCA are (1) extraction of initial components, (2) determination of significant components, retained in a model, (3) rotation of the matrix based on factor loadings to obtain a solution, (4) interpretation of the solution, (5) computation of scores for each factor and of general scores, (6) synthesis of results in a table [27].

In SPSS, the application of this method produces two tables; the first one identifies the principal components and indicates the percentage of variability explained by them, and the second one shows their correlation with the actual variables. The components correspond to the variables to which they are correlated mostly, either positively or negatively, as shown by the value of the correlation coefficient.

In this research, the weights are given by PCA, and adjusted to sum to 100 (i.e., if all principal components explain a percentage of the variability, X, less than 100%, the percentage explained by each of them is increased 100/X times). The approach was used in four case studies. For some of them, additional analyses or data transformations were required.

3. Results

While discussing the results of the research involving the utilization of the method in the four case studies, the purpose of this paper is to reveal the utility of the methodological approach rather than focusing on the relevance of the results, partially published already.

3.1. The level of development in Romania

17 variables (reflecting the foreign direct investments, share of employed population, share of people age 65 and over, the unemployment rate, net earnings per employee, number of inhabitants per room, number of pharmacies, number of physicians, rate of scholar abandon, per capita Gross Domestic Product, number of employees in research, research expenses per person, share of population working in agriculture, use of telephone, use of the Internet, share of modernized roads and infantile mortality) were used to assess the level of development in Romania and map its spatial distribution [28]. The principal components are presented in Table 1, indicating the associations with specific variables.

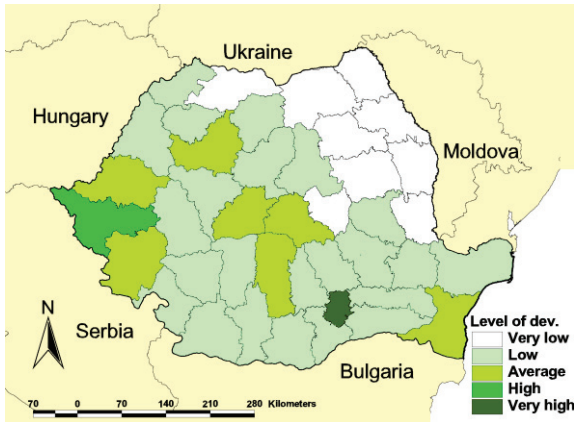

Fig. 1. Showing the spatial distribution of the level of development in Romania determined using Principal Components Analysis-based GIS modeling. A "very high" level of development indicates an economically prosperous region, without social and environmental problems.

Table 1. Results of Principal Component Analysis analyzing the level of development in Romania.

Component	Initial Eigen Value	% variability explained	Corresponding variable	Coefficient of correlation
1	8.59	50.51	Gross domestic product	0.96
2	2.36	13.90	Number of inhabitants per room	-0.78
3	1.21	7.11	Rate of scholar abandon	0.79

PCA found out that out of the seventeen variables only the gross domestic product, number of inhabitants per room and rate of scholar abandon explain almost 72% of the variability. Using their weights, the map displayed in Fig. 2 was obtained through GIS modeling. The map is consistent with other results from the literature [29], indicating Moldova (the eastern part) as the least developed region. Moreover, the three variables describing the level of development belong to different chapter: the gross domestic product is an economic indicator, the number of inhabitants per room is social and rate of scholar abandon describes education.

3.2. The level of development in Ialomița hydrographic basin

15 variables (total population, resident population, livable area per inhabitant, migrations in, migration out, number of people owning a TV, number of pharmacies, number of physicians, population employed in the industry, total employment, total unemployment, employment in the agriculture, active population, population age 65 and over, number of high school graduates) were used to assess the overall level of development in the administrative units of Dâmbovița County situated within the Ialomița hydrographic basin [30]. The principal components are presented in Table 2, indicating the associations with specific variables. The results of PCA indicate that two variables, the number of high school graduates and the population employed in agriculture, account for 84.6% of the total variation, showing that the influence of the major structuring axes represented by communication routes and Ialomița River is crucial, since most settlements close to it exhibit higher levels of development, with anomalies due to the small local influences of small cities – Fieni and Pucioasa [30]. Again, the two variables identified by PCA describe different sectors, respectively education and economy.

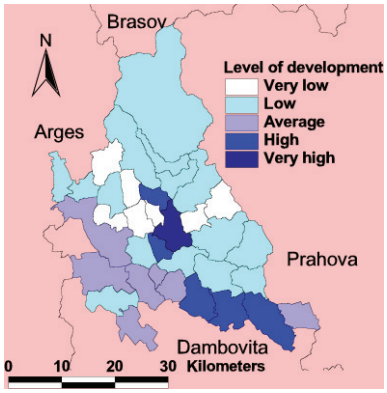

Fig. 2. Showing the spatial distribution of the level of development in Ialomița hydrographic basin determined using Principal Components Analysis-based GIS modeling. A "very high" level of development indicates an economically prosperous region, without social and environmental problems.

Table 2. Results of Principal Component Analysis analyzing the level of development in Ialomița hydrographic basin.

Component	Initial Eigen Value	% variability explained	Corresponding variable	Coefficient of correlation
1	10.39	69.23	Number of high school graduates	0.71
2	1.25	8.30	Population employed in agriculture	0.96

3.3. The development axes of Iaşi County

The study aiming to pinpoint the main development axes within Iaşi County first attempted to map the level of development and then identify the development axes, using 22 variables (the density of population, annual growth, natality, mortality, immigration, emigration, demographic ageing, fertility, dependence of the elderly, per capita gross domestic product, number of companies, unemployment, active population, length of water supply, sewerage and gas supply systems, livable area, number of students per instructor, number of primary and secondary schools, number of high schools, density of higher rank roads and railroads); based on the values for the 98 administrative units, correlated with other qualitative information (literature data and field observations), the development axes were identified and ranked based on existing potentials and perspective for future developments. The principal components are presented in Table 3, indicating the associations with specific variables. The results of PCA indicate that the main variables describing the level of development in Iaşi County are the annual growth, the index of demographic ageing, the employment and migration, relevant to demographic, economic and social issues. The spatial distribution of the level of development based on these variables is illustrated in Fig. 3; it shows that the development occurs at higher rates in the median part of the area, concentrating the most important urban settlements. The next step of the study was to figure out the development axes within the territory of Iaşi County; these are displayed in Fig. 4 [31]. Seven axes with different levels and potentials of development were identified: a main east-west axis, a secondary axis and a tertiary one providing for the western connections of the county, and four potential axes configured in relationship to the evolution of Iaşi City, showing a decreasing trend of the level of development while the distance from the city increases.

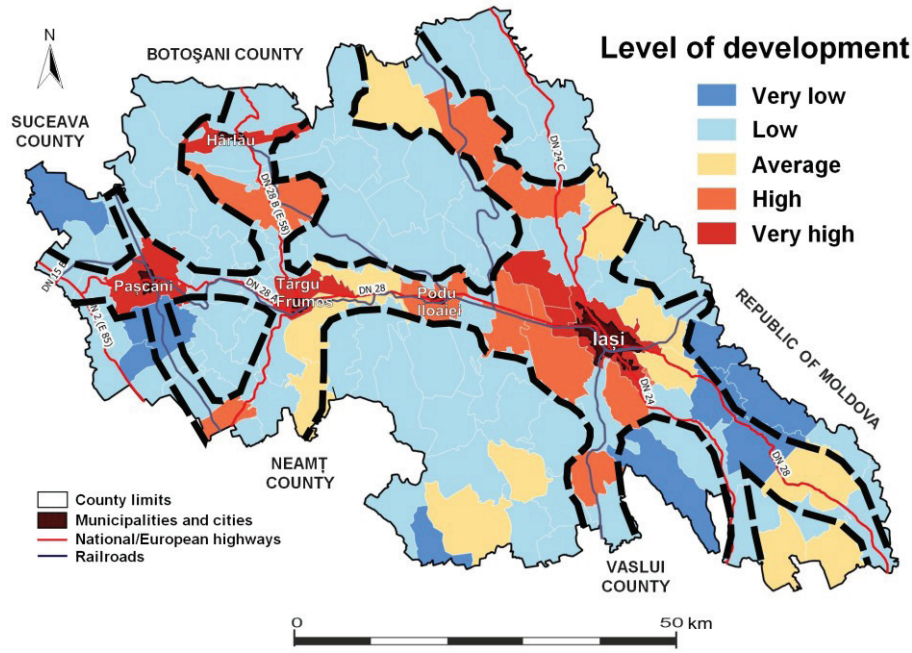

Fig. 3. Showing the spatial distribution of the level of development in Iaşi County determined using Principal Components Analysis-based GIS modeling. A "very high" level of development indicates an economically prosperous region, without social and environmental problems .

Table 3. Results of Principal Component Analysis analyzing the level of development in Iaşi County.

Component	Initial Eigen Value	% variability explained	Corresponding variable	Coefficient of correlation
1	4.35	33.43	Number of employees per 1,000 people	0.91
2	2.50	19.22	Demographic ageing index	0.86
3	1.64	12.61	Migration balance	-0.76
4	1.14	8.75	Average yearly growth	0.72

3.4. Development in a restrictive space: Danube Delta Biosphere Reserve

30 variables (total population, female population, natality, mortality, arrivals, departures, average number of employees in transportation and communication, public administration, commerce, industry, education, finished houses, length of water supply network, total length of the sewerage, livable area, number of PCs, accommodation capacity, number of tourists, people age 65 and over, active population, number of uneducated people, unemployment, number of adobe buildings, number of physicians, population with average and higher education, number of pharmacies, people owning a TV, poverty, income, gross domestic product) were used to determine the principal components used to map the level of development in the administrative units of Danube Delta Reserve of the Biosphere, as a start point for elaborating strategies of development. The principal components are presented in Table 4, indicating the associations with specific variables.

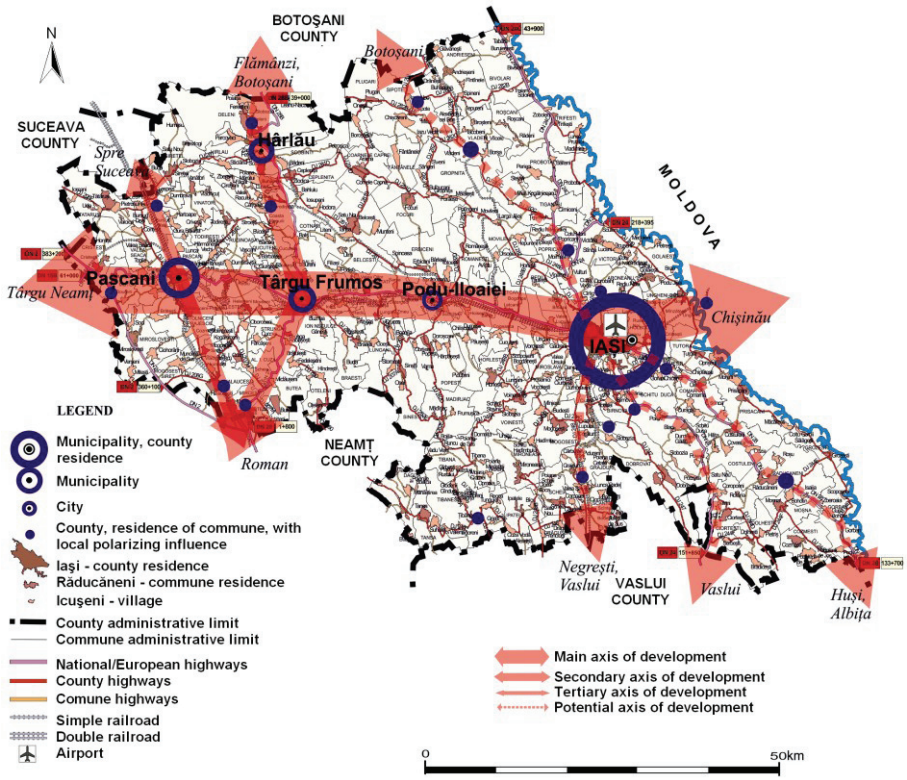

Fig. 4. Showing the spatial distribution of the development axes in Iaşi County determined using Principal Components Analysis-based GIS modeling [31].

PCA showed that the level of development in the administrative units of Danube Delta Biosphere Reserve can be described by the number of departures, number of employees in transportation and communications, number of houses in construction completed and number of employees in commerce, describing social, economic and infrastructure issues. The spatial distribution of the level of development based on these variables is shown in Fig. 5. The results indicate that the main factors restricting the development of the area are accessibility and lack of attractiveness due to the anthropic impact during the communist period, resulting into the transformation of natural areas in agricultural land, nowadays abandoned or undergoing ecological restoration [32].

Table 4. Results of Principal Component Analysis analyzing the level of development in Danube Delta Reserve of the Biosphere.

Component	Initial Eigen Value	% variability explained	Corresponding variable	Coefficient of correlation
1	15.67	52.23	Number of departures	0.99
2	6.26	20.86	Number of employees in transportation	0.93
3	4.26	14.19	Number of employees in transportation	-0.77
4	2.67	8.90	Number of houses completed	0.80
5	1.14	3.81	Number of employees in commerce	0.48

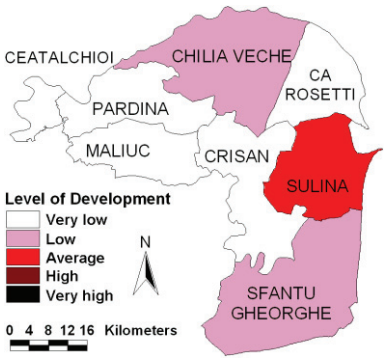

Fig. 5. Showing the spatial distribution of the level of development in the administrative units of Danube Delta Biosphere Reserve determined using Principal Components Analysis-based GIS modeling. A "very high" level of development indicates an economically prosperous region, without social and environmental problems.

4. Discussion

The results shown above can be analyzed from a double perspective. First, from a theoretical standpoint, the research aimed to test the hypotheses according to which development cannot be described from a unique perspective – economic, social, environmental, cultural etc. [33, [34]. In all the cases presented, PCA was applied to assess the level of development in a series of administrative units based on numerous variables describing different aspects: economy, social issues, culture, education, utilities etc., the results constantly show the same – the level of development cannot be assessed by more variables reflecting a single aspect, but only by a set of single variables, each of them reflecting one of the different aspects accounted for. This common characteristic of the results supports the underlying hypothesis.

Second, from a methodological perspective, in each case the results of the method were sound and consistent with the ones of previous studies [29]. This indicates the validity of the approach in addition to their usefulness from a methodological perspective.

Moreover, the third case study points to the potential of the method when used in conjunction with other tools for in-depth research questions, similar to other studies [35]. Simple mapping of the level of development by administrative units was the start point for identifying development axes, based on the pattern revealed using the approach described in this paper.

Nevertheless, there are also limitations of using the methodology, related to the availability of data for smaller administrative unit [36, 37]. In Romania in particular, official statistics are published for each administrative unit only after the censuses carried out at large time intervals, and when such data is made available, only few variables are accounted for.

5. Conclusions

The paper attempted to introduce a methodology based on using PCA in conjunction with GIS modeling to assess the level of development within the territorial subunits of a given region with different sizes, testing the hypothesis according to which the level of development cannot be accurately described from a unique standpoint – economic, social, cultural etc. From a theoretical perspective, the results support the hypothesis. Methodologically, the approach shows, in addition to its utility as a research tool, the potential as a decision-support tool, by pinpointing underdeveloped areas that require a special attention. Nevertheless, its use is limited by the availability of data at the micro-scale.

Acknowledgements

The authors would like to thank anonymous reviewers for their comments that helped improving the quality of the manuscript.

References

[1] Ianoş I, Heller W. *Space, economy, and systems of settlements* [in Romanian]. Bucharest: Technical Press; 2006, p. 6–7.
[2] Ianoş I, Humeau JB. *Theory of the human settlement systems* [in Romanian]. Bucharest: Technical Press; 2000, p. 34.
[3] Rolland-May C. *Evaluation des territoires. Concepts, modeles, methods*. Paris, France: Hermes Science Publications; 2000, p. 27.
[4] Wilson AG. *Complex Spatial Systems: The Modelling Foundations of Urban and Regional Analysis*. Harlow, UK: Pearson Education; 2000; p. 13.
[5] Ianoş I. *Territorial systems. A geographic approach* [in Romanian], Bucharest: Technical Press; 2000, p. 27.
[6] EUROSTAT. *Regions in the European Union Nomenclature of territorial units for statistics NUTS 2006 /EU-27*. Luxembourg: Office for Official Publications of the European Communities; 2007, p. 118–21.
[7] Petrişor AI. Levels of biological diversity: a spatial approach to assessment methods. *Rom Rev Reg Stud* 2008;**4(1)**:41–62.
[8] Petrişor AI. *Systemic theory applied to ecology, geography and spatial planning. Theoretical and methodological developments*. Saarbrücken, Germany: LAMBERT Academic Publishing GmbH & Co. KG; 2011, p. 125–27.
[9] Mehedinţi S. *Terra. Introduction to the science of geography* [in Romanian]. 2nd ed. Bucharest: Encyclopedic press; 1994, p. 149.
[10] Pumain D. An Evolutionary Model of Urban Systems. In: Ianoş I, Pumain D, Racine JB, editors. *Integrated Urban Systems and Sustainability of Urban Life*, Bucharest: Technical Press; 2000, p. 11–34.
[11] Stigler SM. *The History of Statistics. The Measurement of Uncertainty before 1900*. Cambridge, Massachusetts: The Belkmap Press of Harvard University Press; 1986, p. 1.
[12] Vlad I. Reproducibility in computer-intensive sciences. *Ad Astra* 2002;**1(2)**:1–2.
[13] Motulsky H. *Intuitive Biostatistics*. New York: Oxford University Press; 1995, p. 7.
[14] Johnston K, Ver Hoef JM, Krivoruchko K, Lucas N. *Using ArcGIS Geostatistical Analyst*. Redlands, CA: ESRI Press; 2001, p. 143–59.
[15] Petrişor AI. *Elements of regression applied to medical, life and Earth sciences* [in Romanian]. Bucharest: Ars Docendi Press; 2009, p. 7–8.
[16] Ritchie MG, Webb SA, Graves JA, Magurran AE, Macias Garcia C. Patterns of speciation in endemic Mexican Goodeid fish: sexual conflict or early radiation? *J Evol Biol* 2005;**18**:922–9.
[17] Duncan DF. Uses and Misuses of Epidemiology in Assessing Drug Policy. *J Prim Prev* 1997;**17(4)**:375–82.
[18] Bastianoni S, Pulselli FM, Focardi S, Tiezzi EBP, Gramatica P. Correlations and complementarities in data and methods through Principal Components Analysis (PCA) applied to the results of the SPIn-Eco Project. *Journal Environ Manage* 2008;**86(2)**:419–26.
[19] Enăchescu D, Enăchescu C. Principal Components Analysis. Application to the Study of Risk-Factors for Social Dissociation on Territorial Level in Romania. *Austrian Journal of Statistics* 2002;**31(2-3)**:123–30.
[20] Chiţu Z, Şandric I, Mihai B, Săvulescu I. Evaluation of landslide susceptibility using multivariate statistical methods: a case study in the Prahova subcarpathians, Romania. In: Malet JP, Remaitre A, Bogaard T, editors. *Proceedings in Landslide Processes: from geomorphological mapping to dynamic modelling*, Strassbourg, France: CERG Editions; 2009, p. 265–70.
[21] Lefter C, Constantin C. Economic and social disparities of Romania in regional and county profile. *Management & Marketing* 2009;**4(1)**:77–96.
[22] Boamfă I. Mapping of the elements of electoral geography. Case study: Parliamentary elections in Romania and Republic of Moldova after 1989. *Revista Română de Geografie Politică* 2010;**12(2)**:189–205.
[23] Meiţă V, Petrişor AI, Simion-Melinte C-P. Assessing the vulnerability to climate change in the Romanian part of Tisza river basin. *Res J Agric Sci* 2011;**43(3)**:429–436.
[24] Iaţu C, Bulai M. New approaches in evaluating tourism attractiveness in the region of Moldavia (Romania). *Int J Energy Environ* 2011;**5(2)**:165–74.
[25] Mironeasa S, Codină GG, Leahu A, Mironeasa C. Multivariate statistical analysis of Royal Feteasc wine quality from different regions of Romania country. *Food and Environment Safety* 2011;**10(1)**:47–52.
[26] DeCoster J. *Overview of Factor Analysis*. Tuscaloosa, AL: Department of Psychology, University of Alabama; 1998. Retrieved 09/25/2011 from http://www.stat-help.com/factor.pdf
[27] Hatcher L. *A step-by-step approach to using SAS system for factor analysis and structural equation modeling*. Cary, NC: SAS Institute; 1994, p. 1–56.
[28] Ianoş I, Petrişor AI, Zamfir D, Cepoiu AL, Stoica IV. *In search of a relevant indicator measuring territorial disparities in a transition country. Case study: Romania*. *Die Erde* 2012;**143(2)**:in press.

[29] Dobrin M, Tache A, Petrişor AI. System of indicators to analyze regional development disparities in Romania. *Romanian Statistical Review* 2010;**58(8)**:25–37.
[30] Ianoş I, Petrişor AI. Micro-scale geostatistical analysis of the level of development. Case study: mountainous and subcarpathian area of Ialomiţa hydrographic basin. *Geographia Technica* 2010;**Special issue**:47–51.
[31] Iurea D. *Development axes in Iaşi County. Geographical analysis* [in Romanian]. Unpublished doctoral dissertation. Bucharest: University of Bucharest; 2011, p. 127–29.
[32] Văidianu MN. *Designing the development of human settlements in a restrictive space: the Danube Delta* [in Romanian]. Unpublished doctoral dissertation. Bucharest: University of Bucharest; 2011.
[33] Stevens C. Measuring Sustainable Development. *OECD Observer* 2005;**10**:1–8.
[34] Perrons D. Regional performance and inequality: linking economic and social development through a capabilities approach. *Cambridge J Regions Econ Soc* 2011; in press.
[35] Blažek J, Netrdová P. Can development axes be identified by socio-economic variables? The case of Czechia. *Geografie – Sborník ČGS* 2009;**114(4)**:245–62.
[36] Harvey F, Tulloch DL. Local Government Data Sharing: Evaluating the Foundations of Spatial Data Infrastructures. *Int J Geogr Inf Sci* 2009;**20(7)**:743–68.
[37] Loo BPY, Cheng AHT. Are there useful yardsticks of population size and income level for building metro systems? Some worldwide evidence. *Cities* 2010;**27(5)**:299–306.